



Argonne
NATIONAL
LABORATORY

... for a brighter future



U.S. Department
of Energy

UChicago ►
Argonne_{LLC}

A U.S. Department of Energy laboratory
managed by UChicago Argonne, LLC

TRANSIMS Training Course at TRACC

Transportation Research and Analysis Computing Center

Part 3

Population Synthesis Based on CENSUS Resources

Dr.-Ing. Hubert Ley

Transportation Research and Analysis Computing Center

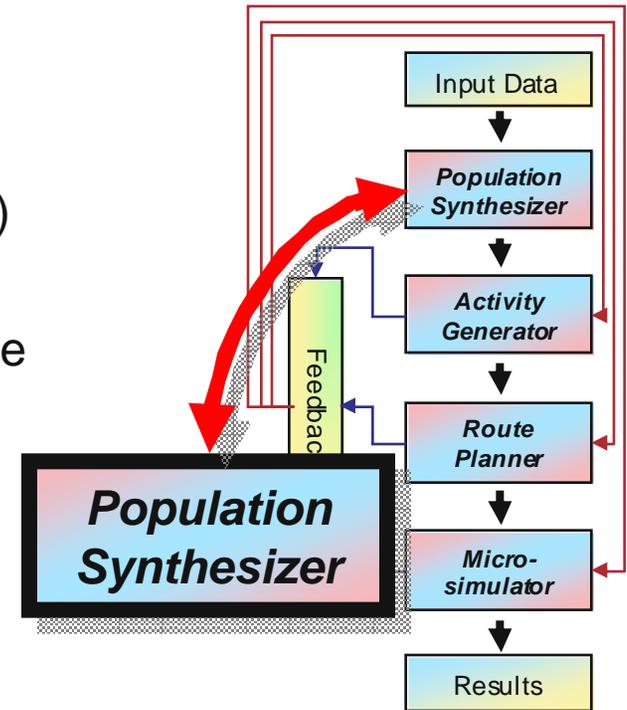
Last Updated: June 12, 2008

Contents

- Introduction
- CENSUS Bureau Data
- Data Sources
- Census Bureau PUMS Data
- Census Bureau STF-3A Data
- Algorithms
- IPF Algorithms (Traditional and Two-Step)
- Three-Dimensional IPF Procedure
- Placement of Households on the Network
- Household Vehicle Ownership

The TRANSIMS Population Synthesizer

- Mimics regional population (“synthetic population”)
 - Demographics closely match real population
 - Households distributed spatially to approximate regional population distribution
 - Household locations determine some of the travel origins and destinations
- Functions of the Population Synthesizer
 - Generation of synthetic households from census data at the block group level
 - Development of each household demographic characteristics (income, members, etc)
 - Placement of each synthetic household on a link in transportation network (activity locations)
 - Assignment of vehicles to each household (sharing vehicles and rides within a household)



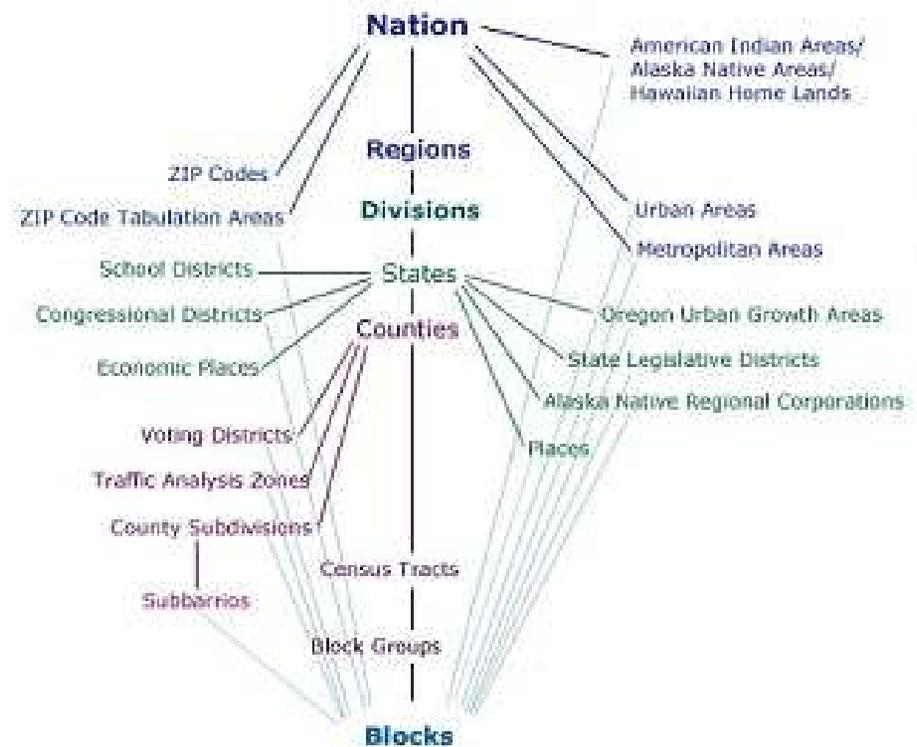
Introduction

- The population generator creates synthetic households from census data to model a real population for the entire coverage area
 - Census data is provided for the entire United States
 - The data is made anonymous by the Census Bureau by
 - *Providing summarized data on various levels (STF)*
 - *Providing small subsets of actual data records taken from a larger area (PUMA)*
 - The smallest unit of public Census data with sufficient detail is at the
 - *block group level (summary data) for ~ 4,000 people*
 - *public use microdata area level (sample records) for ~ 100,000 people*
- Develops associated demographic characteristics for each household
- Places each synthetic household on a link in the transportation network
- Assigns vehicles to each household



Population Synthesizer

- The Census data is broken down by summarizing it on a number of levels:
 - Entire United States
 - By State
 - By County
 - By Census Tract
 - By Census Block Group
 - By Census Block
- Independently, samples are provided for PUMAs (public use microdata areas)
 - 1% and 5% sample records
 - both are made anonymous
- The challenge is to reconstruct a representative synthetic population

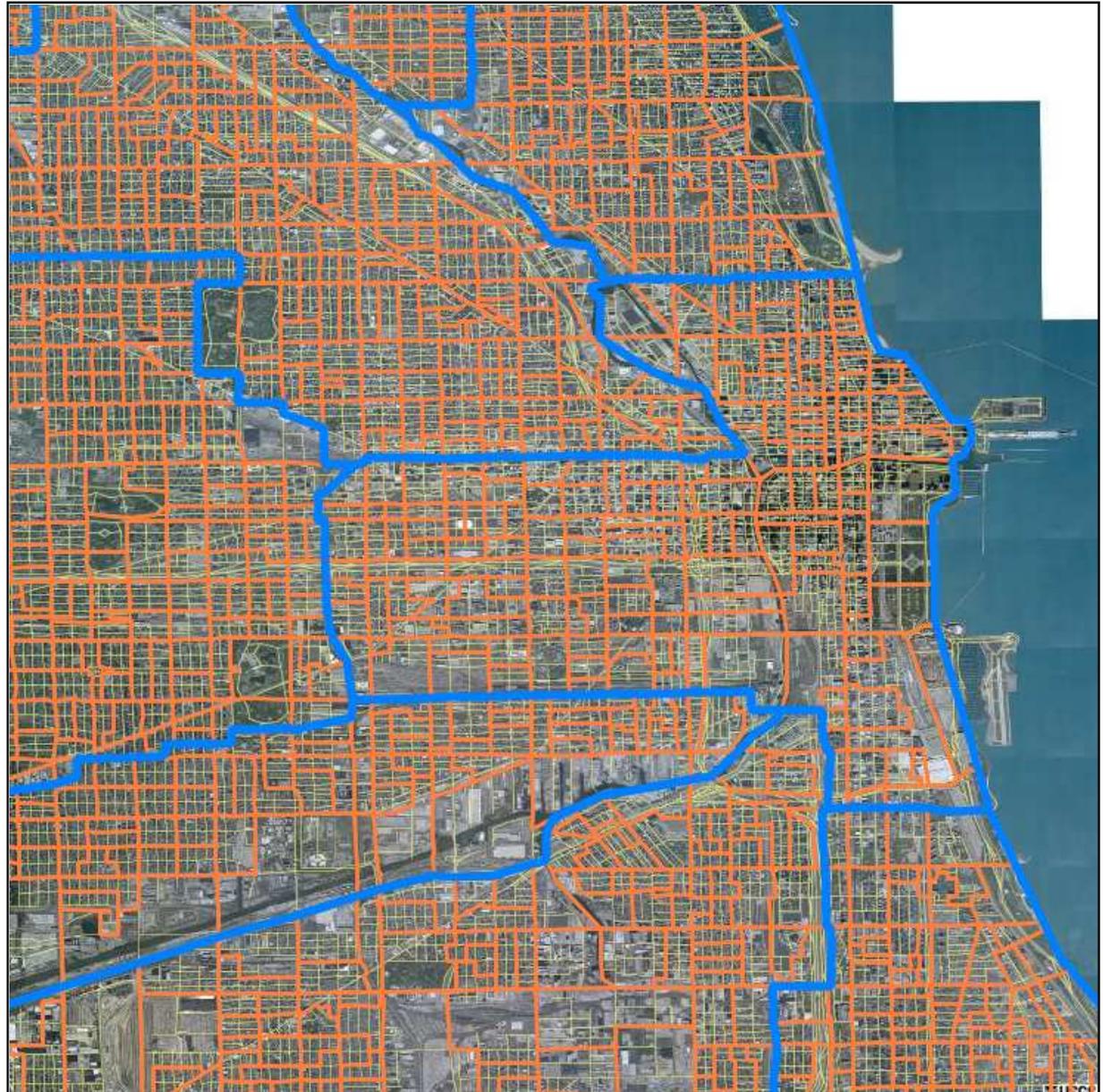
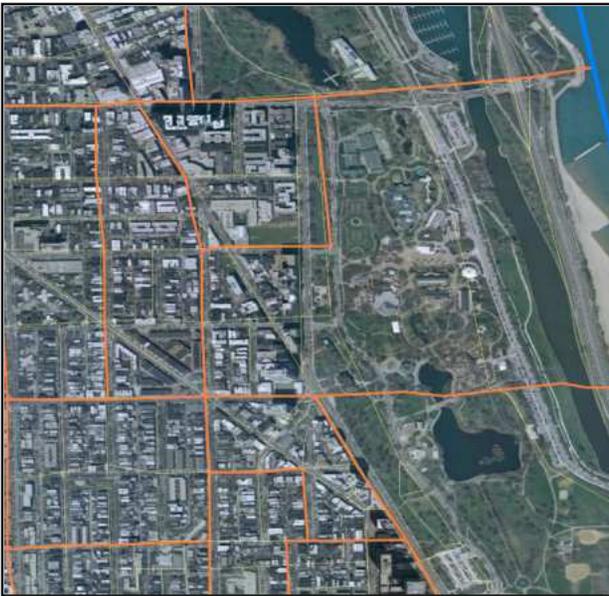


CENSUS Bureau Data

- The following Census Bureau data is of interest for modeling
 - STF-3A tables
 - *Contain demographic summary tables from Census data for small geographic areas. These one-dimensional summary tables contain information on 100% household demographic variables at the Census Block Group level.*
 - PUMS records
 - *Public Use Microdata Sample files consist of a 5% representative sample of almost complete census records from those households contained in a collection of census tracts or other small geographic census areas, which collectively is called a Public Use Micro Area (PUMA).*
- A PUMA is constructed so that it contains approximately 100,000 individuals. These files are edited to protect the confidentiality of all individuals, but they have the information necessary to conduct effective research and analysis.

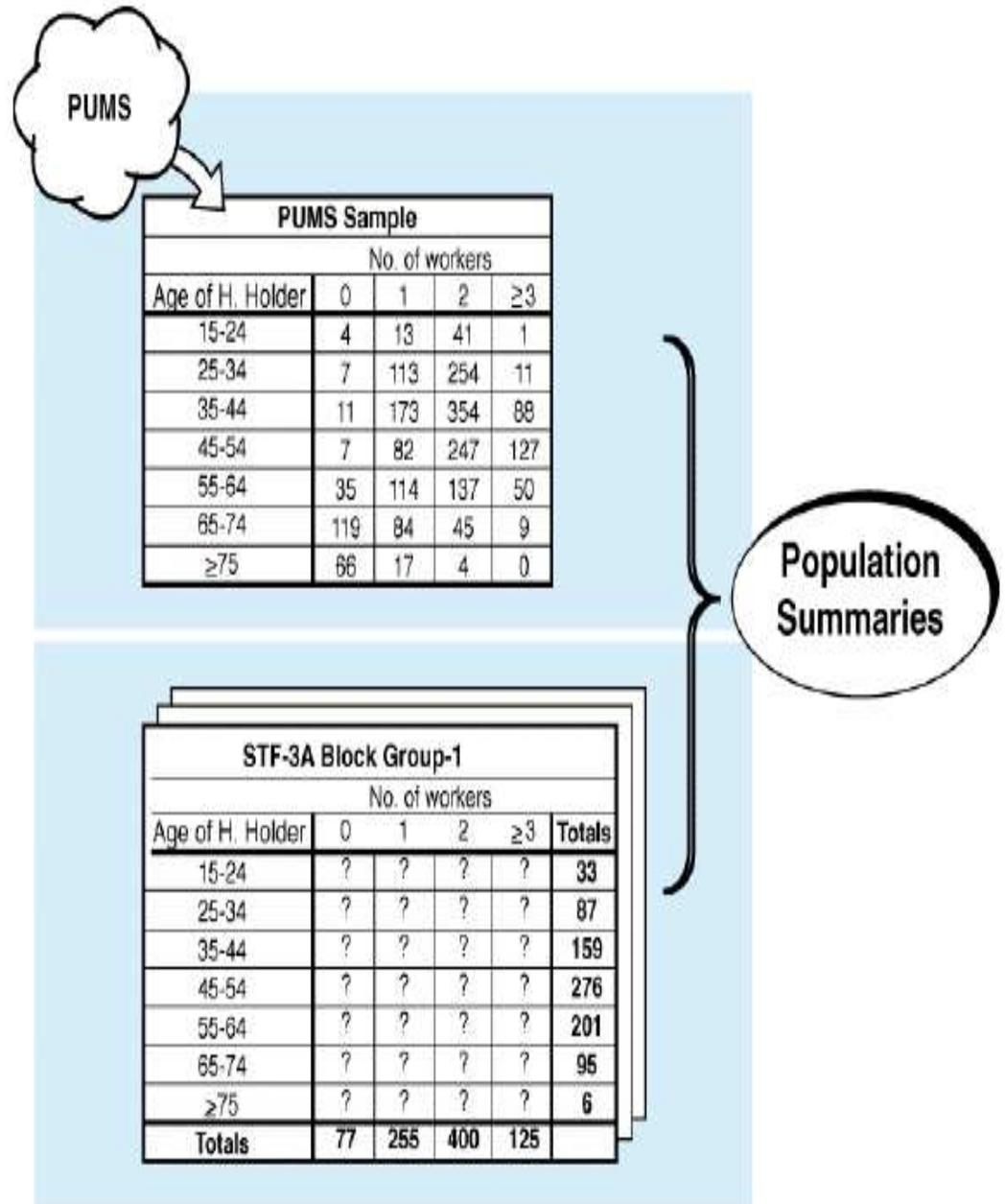
CENSUS Data

- **BLUE**
PUMS 5% Sample Data
- **ORANGE**
Aggregate STF3 Data
- **YELLOW**
Street Network



Data Sources

- The first step in constructing a synthetic population is to use PUMS data and standard census data from STF-3A
- PUMS data contains a 5% sample of real census records modified to maintain anonymity
- These are used in conjunction with STF-3A summary data, which contains marginal demographic information but representing 100% of all households



PopSyn Input Data

- PopSyn is the Version 4 Population Synthesizer
- PopSyn requires the following input data
 - A Zone Data File
 - A PUMS Household File
 - A PUMS Population File
 - An Activity Location File
 - A Process Link File
 - A Vehicle Type Distribution File

- There are two documents that will help understanding the process
 - The 2000 US Census Data Preparation How-To
 - The Population Synthesizer How-To
 - Both are available from the TRANSIMS web site

1.) The Zone Data File

- ZONE, STATE, PUMA are mandatory fields
- The TOTAL field contains the total number of households present in each zone
- The ZONE field represents any geographic boundary, be it census tracts, block groups, traffic analysis zones, blocks, or any user-specified area for which household totals are available and for which household attributes need to be synthesized
- Two control attributes are used in this file, namely household size HHSIZE and auto ownership AUTOS to guide the population synthesis process
- Two categories (HHSIZE and AUTOS) in the example below

ZONE	STATE	PUMA	TOTAL	HHSIZE1	HHSIZE2	HHSIZE3	AUTOS1	AUTOS2
1	DC	200	400	0.25	0.4	0.35	0.6	0.4
2	DC	200	600	0.28	0.46	0.26	0.7	0.3
3	VA	100	400	0.25	0.4	0.35	0.6	0.4
4	VA	100	400	0.28	0.46	0.26	0.7	0.3
5	VA	100	200	0.26	0.46	0.28	0.5	0.5

2.) The PUMS Household File

- This file contains a representative sample of household records with complete information
- WEIGHT is being normalized and does not need to be a percentage
- The HHOLD, STATE, and PUMA fields are required
- Alternative names for HHOLD are HOUSEHOLD, HH_ID, HHID, or HH
- Additional attributes may be included:
 - They will not have any influence on the synthesis of household demographics
 - however, they will be incorporated in synthesized households in the output household file)

HHOLD	STATE	PUMA	WEIGHT	HHSIZE	AUTOS	
1	DC	200	15	1	2	
2	DC	200	25	2	1	
3	DC	200	11	3	1	
4	DC	200	50	3	2	
5	AZ	600	150	1	2	
6	AZ	600	205	2	1	
7	AZ	600	101	3	1	

3.) *The PUMS Population File*

- This file includes the person characteristics for each household included in the PUMS household file (which represents e.g. a 5% sample of actual records)
- For example, person 1 in household 2 is a 30 year old (AGE = 30) male (GENDER = 1) worker (WORK = 1) who drives (DRIVE = 1)
- The records do not influence the population synthesis process
- Every synthesized household inherits these population/person records

	HHOLD	PERSON	AGE	GENDER	WORK	DRIVE
	1	1	30	1	1	1
	2	1	20	2	1	1
	2	2	10	2	0	0
	3	1	40	1	1	1
	3	2	38	2	1	1
	3	3	6	1	0	0
	4	1	25	2	1	1
	4	2	8	1	0	0
	4	3	2	2	0	0
	5	1	30	1	1	1
	6	1	20	2	1	1
	6	2	10	2	0	0
	7	1	40	1	1	1
	7	2	38	2	1	1
	7	3	6	1	0	0

4.) The Activity Location File

- This is a TRANSIMS network file, typically generated by TransimsNet
- Additional columns are needed to indicate membership to a geographic entity
 - TAZ (Traffic analysis zone, provided to and allocated by TransimsNet)
 - TRACT (Census Tract)
 - BG (Census Block Group)
 - USER1 (a weight field for the distribution when multiple locations are allocated within the same zone)

ID	NODE	LINK	OFFSET	LAYER	EASTING	NORTHING	ELEVATION	TAZ	TRACT	BG	AREA	USER1
1	5	1	666.7	WALK	1985	4666.7	0	1	1	1	3	6
2	1	1	333.3	WALK	2015	4666.7	0	2	1	2	3	2
3	5	1	333.3	WALK	1985	4333.3	0	1	2	1	3	6
4	1	1	666.7	WALK	2015	4333.3	0	2	2	2	3	2
5	6	2	666.7	WALK	2985	4666.7	0	2	2	3	3	8
6	2	2	333.3	WALK	3015	4666.7	0	3	3	1	3	4
7	6	2	333.3	WALK	2985	4333.3	0	2	3	2	3	8
8	2	2	666.7	WALK	3015	4333.3	0	3	3	3	3	4
9	7	3	666.7	WALK	3985	4666.7	0	3	4	1	3	10
10	3	3	333.3	WALK	4015	4666.7	0	4	4	2	3	6

5.) The Process Link File

- This is a TRANSIMS network file, typically generated by TransimsNet
- The process link file assigns a delay and cost for each movement from an activity location, transit stop, or parking location to another one of those
- There is probably no need to manually edit the file

ID	FROMID	FROMTYPE	TOID	TOTYPE	DELAY	COST	NOTES
1	1	ACTIVITY	1	PARKING	30.0	0.0	Parking Access
2	1	PARKING	1	ACTIVITY	30.0	0.0	Parking Access
3	2	ACTIVITY	2	PARKING	30.0	0.0	Parking Access
4	2	PARKING	2	ACTIVITY	30.0	0.0	Parking Access
5	3	ACTIVITY	3	PARKING	30.0	0.0	Parking Access
6	3	PARKING	3	ACTIVITY	30.0	0.0	Parking Access
7	4	ACTIVITY	4	PARKING	30.0	0.0	Parking Access
8	4	PARKING	4	ACTIVITY	30.0	0.0	Parking Access
9	5	ACTIVITY	5	PARKING	30.0	0.0	Parking Access
10	5	PARKING	5	ACTIVITY	30.0	0.0	Parking Access

6.) *The Vehicle Type Distribution File*

- The vehicle type distribution file provides the distribution of the vehicle fleet within the project area.
- When a vehicle is generated, it is assigned a type based on this distribution (SHARE values)

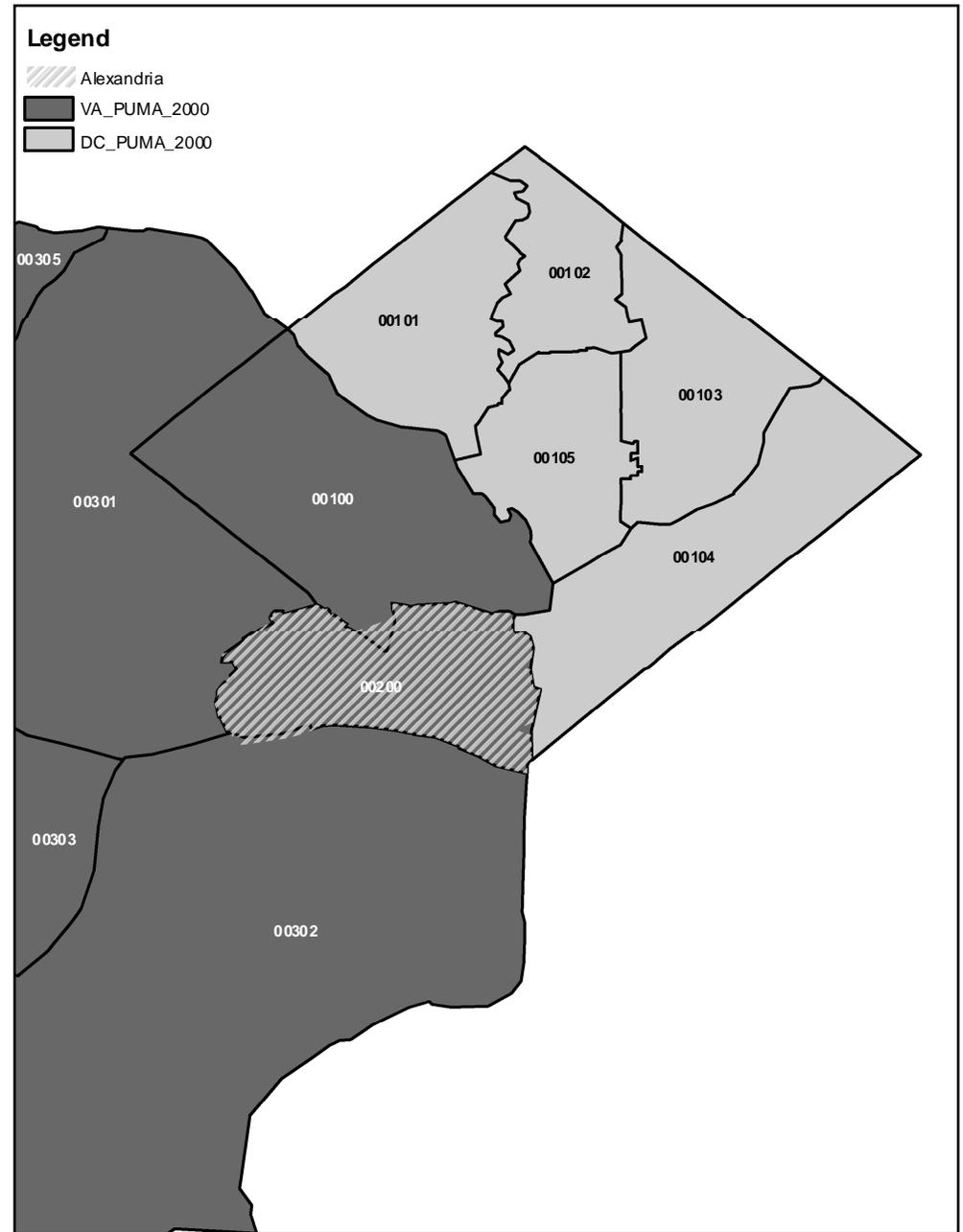
TYPE		SUBTYPE	SHARE
1	0	50	
1	1	50	
2	0	20	

Data Preparation

- Identify Appropriate PUMAs
- Associate Block Groups with a PUMA
- Associate Activity Locations with a Block Group
- Prepare STF-3A Files
- Prepare PUMS Data

Identify Appropriate PUMAs

- Download the PUMA shape files (boundary files for any state in the US may be downloaded from http://www.census.gov/geo/www/cob/bdy_files.html)
- Compare the PUMA shapes with your simulation area in uDig or ArcMap
- Ensure that you are using appropriate projections so that the data superimposes accurately



Associate Block Groups with a PUMA

- In **ArcMap**, this can be done as follows:
 - Download block groups shape files
 - For VA as: <http://www.census.gov/geo/www/cob/bg2000.html#shp> → bg51_d00_shp.zip
 - The coordinate system for this file is GCS_North_American_1983
 - Unzip the shape files and load them into **ArcMap**
 - Use the Join option to tag block groups with PUMAs
 - The result looks like this:

- An alternative is the use of Mable/GeoCorr, a method that was used in previous TRANSIMS versions.
 - Details in the How-To Documents

FID	Shape	FID_1	STATE	COUNTY	TRACT	BLKGROUP	PUMA5
0	Polygon	0	51	069	0503	1	00400
1	Polygon	1	51	069	0503	2	00400
2	Polygon	2	51	069	0503	3	00400
3	Polygon	3	51	069	0504	1	00400
4	Polygon	4	51	069	0502	2	00400
5	Polygon	6	51	069	0501	1	00400
6	Polygon	9	51	069	0502	3	00400
7	Polygon	10	51	069	0501	2	00400
8	Polygon	11	51	069	0502	1	00400
9	Polygon	13	51	069	0505	3	00400
10	Polygon	16	51	069	0501	4	00400

Associate Activity Locations with a Block Group

- TransimsNet creates the original activity locations file
- Create a GIS shape file from the activity locations file using ArcNet
- Load the block group shape files into ArcMap
- Use the Join option to associate each activity location with the block group it falls within
- There is some cleaning of the activity locations file necessary, described in the How-To



Prepare STF-3A Files

- Before designing the Population Synthesis model, the user must consult the type of information collected in the STF-3A files to make sure relevant data is available.
- The list of summary tables available in the STF can be found in
 - http://www2.census.gov/census_2000/datasets/Summary_File_3/0SF3_table_matrix.doc
- The following demographics are selected to control the population synthesis for Alexandria, namely
 - The age of householder (STF-3A Segment 01, Table P13),
 - Household size (STF-3A Segment 01, Table P14), and
 - Household income (family: STF-3A Segment 07, Table P76; non-family STF-3A Segment 07, Table P79)
- The appropriate ZIP files can be downloaded from the Census site
- There is also a Microsoft Access Database for download that should be used to process the data files

Prepare STF-3A Files

- The complete sample procedure for Alexandria is described in the How-To document
- The resulting files for use by PopSyn should look like this:

SUMLEV	LOGRECNO	STATE	COUNTY	TRACT	BLKGRP	UNIQUEID	PUMA	TOTAL
HHAGE1	HHAGE2	HHAGE3						
090	0001067	VA	013	100100 1	510131001001	100 1995	43 50	46
090	0001068	VA	013	100100 2	510131001002	100 1990	12 42	50
090	0001069	VA	013	100100 3	510131001003	100 2024	40 65	35
090	0001070	VA	013	100100 4	510131001004	100 2160	13 73	82
090	0001071	VA	013	100100 5	510131001005	100 1970	24 45	65
090	0001072	VA	013	100100 6	510131001006	100 1170	8 39	26
090	0001075	VA	013	100200 1	510131002001	100 2170	8 59	29
090	0001076	VA	013	100200 2	510131002002	100 2380	21 39	60
090	0001077	VA	013	100200 3	510131002003	100 1740	11 40	55
090	0001078	VA	013	100200 4	510131002004	100 2910	29 60	14

Prepare PUMS Data

- The PUMS file contains two types of records: household records and person records.
- The household records start with an “H” identifier whereas the person records start with a “P” identifier.
- Both records are linked based on a SERIALNO, which will be used in this demonstration in lieu of the household number.
- The file is in fixed text format, and as such requires a small script to extract the data into appropriate fields

```
H000134755135001005101188728840887288405153 67246218 67002719 229668597
P000134701000023010002203201101000001864010203001000110226999011062503031205019980203
P000134702000017030110201001101000001864010500002030020226999011062502031205019980203
H000436655135001005101188728840887288405153 67246218 67002719 229668597
P00043660100003101000010440010110000010147050600100013014205002200000 002701000000100
P00043660200002019000010500010110000010147050600100014014812902200000 005601000000100
H000584755135001005101188728840887288405153 67246218 67002719 229668597
P000584701000034010000101800102100001965480506002060100997999011077702024905019990202
P000584702000034160000102400102100001965480506002060110997999011077703024905019970202
H000764655135001005101188728840887288405153 67246218 67002719 229668597
P00076460100001101000010280010110000010147050600307015005199901200000 002611000000300
P00076460200001118000010280010110000010147010200307015008899901200000 002601000000300
P00076460300000818000420260020100000186401010200307014021099901200000 004801000000300
```

Census Bureau PUMS Data

- Typical information that is being extracted and used from the PUMS data sets

Data	Record Size	Beginning At Field	Description of Data	Allowed Values	Description of Values
RECTYPE	1	1	Record Type	H	Housing Record
PUMA	5	13	Public use microdata area (state dependent)	00100... 99999	PUMA code.
RHHINC	7	141	Household income	0000000	N/A* (GQ**/vacant/no income)
				-999999... 9999999	Total household income in dollars
RWRKR89	1	148	Workers in family in 1989	0	N/A*
				1	No Workers
				2	1 Workers
				3	2 Workers
R18UNDR	1	162	Presence of person under 18 years in household	0	N/A* (No person under 18 in household/GQ**/vacant)
				1	1 or more person under 18 in household

Data	Record Size	Beginning At Field	Description of Data	Allowed Values	Description of Values
RECTYPE	1	1	Record Type	P	Person Record
RELAT1	2	9	Relationship	00	Householder
				01	Husband/Wife
				02	Son/Daughter
				03	Stepson/Stepdaughter
				04	Brother/Sister
				05	Father/Mother
				06	Grandchild
				07	Other relative
				Non Related	
				08	Roomer/boarder/foster child
				09	Housemate/roommate
				10	Unmarried partner
				11	Other non-relative
Group Quarters					
12	Institutionalized person				
13	Other person in group quarters.				
SEX	1	11	Sex	0	Male
AGE	2	15	Age	1	Female
				00	Less than 1 year
				01... 89	Age in years
				90	90 or more years old
WORK89	1	122	Worked last year (1989)	0	N/A (less than 16 years old)
				1	Worked last year
				2	Did not work last year

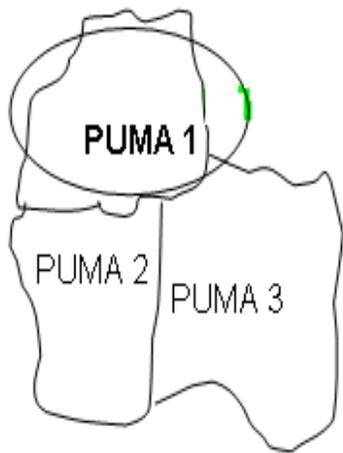
Map of Block Groups and Census Tracts for Chicago

- Synthetic households in TRANSIMS are divided into three categories:
 1. Family households — two or more related persons
 2. Non-family households — persons living alone or unrelated persons living together
 3. Group quarters — dwellings such as prisons or college dormitories

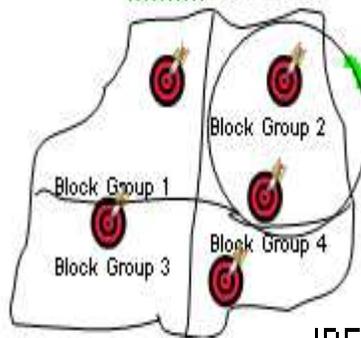


Relationships Among PUMA, PUMS, Block Group, and STF-3A

The Study Area comprised of 3 PUMA



PUMA 1 comprised of 4 block groups



IPF uses PUMS data to fill the unknown cells in STF-3A

STF-3A for Block Group 2

?	?	?	<input checked="" type="checkbox"/>
?	?	?	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	



<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	

PUMS of PUMA 1

= Sampled Households, PUMS (5% of PUMA)

= Known Cross-Tabulated Demographic Variables

IPF Algorithms (Traditional and Two-Step)

- The algorithm adopted by the TRANSIMS Population Synthesizer is based on two types of IPF algorithms:
 - The traditional IPF procedure proposed by Deming and Stephan (1940).
 - The two-step IPF procedure developed by Beckman (1996), known as the modified IPF procedure .
- Traditional Procedure fits only one block group at a time.
- Two-step Procedure can simultaneously consider all block groups that make up a PUMA.
- The two-step procedure makes use of the traditional IPF procedure in its analysis.

Three-Dimensional IPF Procedure

- The same algorithm can be extended for additional dimensions
- The following is a typical example

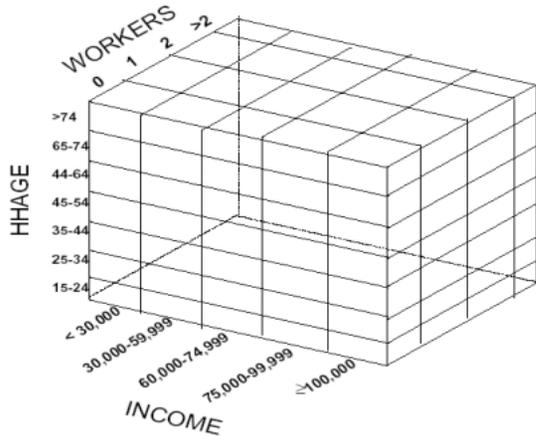
- Age of Householder
 - $15 > \text{HHAGE} < 24$
 - $25 < \text{HHAGE} < 34$
 - $35 < \text{HHAGE} < 44$
 - $45 < \text{HHAGE} < 54$
 - $55 < \text{HHAGE} < 64$
 - $65 < \text{HHAGE} < 74$
 - $\text{HHAGE} > 74$

- Household Income
 - $\text{INC} < 30,000$
 - $30,000 < \text{INC} < 60,000$
 - $60,000 < \text{INC} < 75,000$
 - $75,000 < \text{INC} < 100,000$
 - $\text{INC} > 100,000$

- Number of Workers
 - $\text{WORKERS} = 0$
 - $\text{WORKERS} = 1$
 - $\text{WORKERS} = 2$
 - $\text{WORKERS} > 2$

Three-Dimensional IPF Procedure

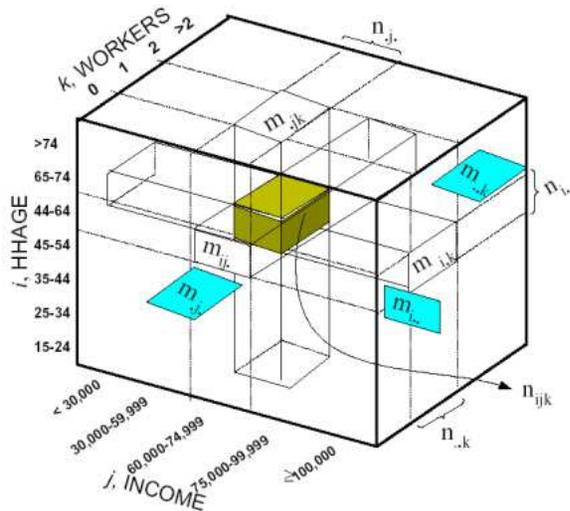
The generalized mathematical formulation developed by Deming *et al.* [1940] for the 3-dimensional matrix of $7 \times 5 \times 4$ can be stated as follows:



$$m_{ijk}' = n_{ijk} \frac{m_{i..}}{n_{i..}} \tag{I}$$

$$m_{ijk}'' = m_{ijk}' \frac{m_{.j.}}{m_{.j.}'} \tag{II}$$

$$m_{ijk}''' = m_{ijk}'' \frac{m_{..k}}{m_{..k}''} \tag{III}$$



- n = In general, it refers to a cell or a marginal value in the PUMS data,
- m = In general, it refers to the marginal value of the STF-3A matrix,
- n_{ijk} = Sample frequency from PUMS data falling in the cell n_{ijk} (please see the figure below for illustration),
- $m_{.j.}'$ = Marginal data of the updated matrix (from PUMS) for the second (j) dimension,
- $m_{..k}$ = Marginal data of the STF-3A file for the third (k) dimension, for example here the third (k) dimension is the WORKERS variable, and
- $m_{..k}'$ = Marginal data of the updated matrix for the third (k) dimension.
- $m_{..k}$ = Marginal data of the STF-3A file for the third (k) dimension, WORKER
- $m_{.j.}'$ = Marginal data of the updated matrix (from PUMS) for the second (j) dimension,
- $m_{..k}''$ = Marginal data of the updated matrix for the third (k) dimension.

Two-Step (Modified) IPF

- TRANSIMS uses a two-step IPF procedure instead
- Traditional procedure fits only one block group at a time
- Beckman showed that fitting only one block group at a time may not be entirely correct
 - The sum of the block's STF-3A should also have the same correlation structure as the PUMS data, which equally represents all the blocks in a PUMA
- The two-step IPF procedure can simultaneously consider all block groups that make up the PUMA
- Details on the Two-Step IPF can be found in the TRANSIMS documentation, in particular the
 - The 2000 US CENSUS Data Preparation How-To
 - The Population Synthesizer How-To

PopSyn in a Nutshell

- Define the household attributes and classification for each PUMA as specified by the user
- Construct the zone marginal totals (summaries) for each identified attribute
- Construct the PUMS cross-classification (sample PUMA cross-classification to be more precise) using the same attributes and classification system specified by user
- Aggregate all marginal tables for all zones within a PUMA
- Apply an IPF process to estimate the cross-classification table for each PUMA using the aggregate PUMA marginal's table and the PUMS cross-classification table
- Apply an IPF process to estimate the zonal cross-classification tables using disaggregate marginal tables and the estimated PUMA cross-classification table as an additional marginal table and a PUMS-like cross-classification table consisting entirely of ones.
- Select a PUMS household for each household in the zone cross classification and randomly locate it to a weighted activity location within the zone
- Randomly select a vehicle type for each vehicle assigned to the household and locate it at the parking lot attached to the activity location and output the vehicle record

Credits and Acknowledgements

- GIS visualization materials were mostly developed at Argonne based on the TRANSIMS tools developed by AECOM for USDOT
- Chicago road and transit network data used in some of the examples was provided by the Chicago Metropolitan Agency for Planning
- USDOT provided the funding for the development of these training materials
- USDOT provided the funding for the TRACC computing center and the resources necessary to perform these training session
- Some materials have been developed for USDOT by Prof. Antoine Hobeika, Virginia Polytechnic Institute, Civil and Environmental Engineering
- The presentation is based on materials provided by USDOT at a training course in November 2006, and has been updated using the information available in the How-To documentation from the TRANSIMS web site